Relational Databases at the S.C. Geological Survey

Process steps for developing a data dictionary in preparation of a drill-hole relational database

C. Scott Howard
Department of Natural Resources, Geological Survey
February 2008

Table of Contents

List of Figures

Appendices

"Somebody has to do something, and it's just incredibly pathetic that it has to be us"

Jerry Garcia describing the Grateful Dead's participation in a benefit concert to help preserve the Amazonian rain forests.

Problem Statement

The South Carolina Department of Natural Resources (DNR) is in the developmental stages of establishing relational databases[1] for various programs and sections. The Land, Water, and Conservation Division (LWC) and the Geological Survey (SCGS) have made the creation of electronic databases a priority through their operational plans. There are numerous scientific data sets within DNR, and for the most part they are inaccessible or underutilized by a majority of data users. A robust digital database of natural resources has several benefits: organization of data (we know what we have); development of baseline data sets (documentation of spatial and temporal change); assessment of data quality (we know how good our data are); easy access to data input and retrieval (prevention of institutional data loss). There is another significant advantage to a database. A robust, organic database is an invaluable research and decision-making tool. The ability to access large natural-resource data sets is essential. It allows users to answer questions that are constantly being put before them as part of DNR's mission goals.

To date, the Geological Survey has not transferred any of its data collections to a digital database. The Geological Survey is extremely interested in developing databases for drill logs, core samples, publications, and geologic maps. In addition to fulfilling current operational plans for database development, the advent of GIS in land-use planning and economic development necessitates that the Geological Survey deliver geologic information in suitable electronic formats. Although it is beyond the scope of this project to design and build a relational

[1] Bold-faced words are included in the glossary.

database, it is possible to lay the groundwork by establishing a framework for developing a database.

The purpose of this project is to model front-end processes that contribute to the end goal of developing a relational database. There are two things needed before database design and programming begin: a data dictionary and a data model. A data dictionary defines the individual pieces that make up the data, and a data model describes the way individual pieces relate to one another. The first goal of this project is to develop a data dictionary of the Geological Survey's drill-hole data set. The second goal is to model this process so that it can be applied to any data set within DNR. In other words, develop a universal process that can help others develop their own data dictionary and, ultimately, an operational relational database. As an addendum to the first goal, portions of this paper are devoted to enumerating some of the processes and problems to be expected once the data dictionary is complete. This results in two secondary goals of this project. First, ideas and process models are explored and explained on how to validate data, how to handle it during data entry, how to avoid or anticipate problems during development of the database, and how to use standards to ensure data quality and accuracy. The final step of this project is more precisely a commitment. Because of the Geological Survey's acute desire to acquire a drill-hole database, the project cannot end with the submission of this paper, and so a commitment is needed to see that the initial steps proposed by this project are followed to their end goal, a drill-hole database.

As a further preamble, the assignment of work efforts in this project is addressed. An initial reaction might question why a non-database specialist would be involved in assembling a

data dictionary. Why not leave it up to the database experts? There are several motivations for data users and data producers to work on developing a data dictionary. First, who better to define and categorize data than the originators of the data? An intimate understanding of the data, how it's collected, and how it is used is a significant advantage. This is a primary reason for enlisting data users rather than data modelers in this process. A second reason is that by allowing data producers and users to critique and analyze their own data and recording methods, you can also initiate the development of data standards. The process of developing a data dictionary leads to an introspection on the nature of your data, which allows you to document accurately and completely your collection and recording methods and to describe issues of data accuracy and quality. The answers to these fundamental questions about data result in each data element being clearly defined and distinguished from other elements. These answers can also form the basis for developing data standards. Why data standards? Data standards are important because they ensure consistency and accuracy. You know how the data is collected; therefore you trust the accuracy and quality of the data. Data standards are a way of formalizing data techniques. Measurements are collected in specific formats, taken at specified time intervals, taken with specific instruments, and measured to specified levels of accuracy. This insures consistency, which, in turn, addresses accuracy and quality issues. The reasons for collecting and recording high quality data are numerous. For example, for natural resource data to be of use, particularly in a GIS environment, data records (see data hierarchy and Figure 1) must include some mandatory data elements and minimum requirements. If a data record lacks some of these required elements, the entire data record could be less than useful to the point of deleting it from the data set. Again, because the majority of natural-

resource data is fixed by geographic coordinates, accurate locations matter. One known problem with old data sets in DNR is that data locations are sometimes non-existent. Any associated data with that record is, therefore, lost, unless the location can be reconstructed, which is problematic in itself. Another reason for standards is to ensure the viability of your data in the future. Standards would ensure that old data and its formats conform to present and future data formats, thus assuring continuity of the database.

Another aspect of data evaluation is predicting how the database might change in the future, or more likely, how it will evolve and grow. The database should be designed to handle expanding data needs and changes in data collection in the future. To a certain degree, data evolution can be predicted. A thorough comprehension of your data holdings and collection techniques combined with future plans can give insight into future data needs. These needs would include your expectations, as well as your customers.

Data Collection

For this project, data collection is a relatively minor component. The majority of information needed to complete the project tasks has been collected over years of practical work experience. The emphasis of project work is in the analysis and application of those results towards understanding how a database can be established. Nevertheless, the primary goal of data collection was to obtain a fundamental understanding of the data hierarchy of the Geological Survey's drill-hole data collection, and translate this information into a data dictionary. Gathering information about drill-hole records (see Appendix 1) involves reviewing and assimilating the existing files in order to identify natural data elements, data fields, data

records, and data tables. First-hand experience with drilling and mapping projects and writing descriptive lithologic logs also contributed to a thorough understanding of the data collected. Additional information about data sets and collection methods were gathered from personal interviews of SCGS geologists. There are over 6000 drill-hole records in SCGS files (Ralph Willoughby, oral communication, 2007). These records are from various drilling projects in the Coastal Plain of S.C. during the last 40 years. They contain valuable lithologic and stratigraphic information that needs to be captured in a digital format. In addition to existing information in drill records, the database will need to handle future data entries, which may not be similar to the old data standards or format. Therefore, the database needs to flexible to handle new types of data, which is characteristic of relational databases. For example, some new areas of data collection may involve ground-water measurements, alternate stratigraphic picks, detailed textural studies of sediments, comments about fossil or mineral content that identify the source or age of the unit, and links to other data sources such as maps, geophysical logs, cross sections, or pictures. These other data sources constitute separate databases that will eventually need to be constructed. A positive outcome of this project might lead to the development of these new databases.

Data Analysis

While assembling and evaluating the data used to construct the data dictionary (Appendix 2), several key findings were made that merit discussion, including observations on completeness of the data set, data quality and accuracy, a model for transferring data to digital format, and a final evaluation of developing data standards.

The premise that the people who work with geologic data have the greatest insight is now clear. Therefore, to develop a dictionary that describes and defines elements of geologic drill-holes, go the source. On the basis of discussions with SCGS geologists three major data types were identified: field data (primary information collected on site), monitoring data (long term data collection at a site), and research or refined data (in-depth analyses of samples developed off site, e.g. chemical analyses, sediment analyses, age data, and fossil and mineral studies). The Geological Survey collects primarily field data and research data. The drill logs for the database are classified as field data.

Further discussions found that data records of any type contain three basic elements. For any data record to be of use, it must contain these three elements: identification label, data location, and lithologic log. If any element is missing, the data record is incomplete, and its value to the database is questionable. If the record cannot go into the database, should it be discarded? Because this is a permanent deletion of information, this choice should be a last resort. Alternatively, it might be possible to rehabilitate a bad record at some later time.

A wide range in data accuracy was found. Accuracy of data is principally associated with locational information, which is the single most important element of a data record. A unique site identifier is usually present, and SCGS has developed an internal system for assigning ID numbers to drill sites. Because SCGS geologists have generated detailed lithologic logs for each drill site, the lithologic log is assumed to be complete and accurate.

To assist in data evaluation, a flow-chart analysis was constructed to model the data input process (Figure 2). The flow chart is designed with drill records in mind, but it is general

enough that it could accommodate most other data sets (Figure 3). The flow charts demonstrate a work-flow process to quickly evaluate data prior to database entry. It allows a non-experienced user to make decisions about data without involving immediate supervision. As a subsidiary to the flow chart, a cause and effect diagram (Figure 4) models the problems associated with bad data elements. The primary causes of bad records are transcription errors, omissions during original data recording, and old records.

Transcription errors are mostly attributed to keystroke errors during data entry, but misreading and bad handwriting are other sources. The data originator can cause error through neglecting to record all the pertinent information at a location. This problem is significant because with multiple geologists producing drill-hole logs, each may have their own particular logging scheme. A solution is to develop a standardized logging form that requires specific data to be collected. A standardized data form would be a reminder to the geologist of the data that is necessary. Another source of bad data records is a correlation of record age to completeness and accuracy. The older a record is the greater the chance that the location is missing or inaccurate, or it may not have a complete identity label, or the lithologic log is missing or incomplete. This is almost entirely due to data needs and uses. In the early days of the Survey, data was gathered almost entirely for economic applications and, therefore, site specific. They were focused on the detail rather than assembling a statewide data set. Although they valued the data they collected for projects, their data needs did not correspond to our present-day requirements in today's digital environment.

The assembly and evaluation of data is an effort that provides greater insight into the data; it is an introspection process that gives greater meaning to the value and necessity of the data. This introspection can be turned into another process of developing data standards. How do we ensure that the data we have is good? How do we make sure that the data we collect in the future is of high quality? By knowing the characteristics of the components of a data record more thoroughly, it is possible to establish requirements for the data collected.

Part of the effort in the database development is evaluating the quality of data going into the database, separating the wheat from the chaff. The first cut off for drill-hole data is assessing whether a record contains the 3 essential data elements. On paper records, this information is divided into header material (ID label and site location) and the written lithologic log (Appendix 1). Header data consists of information about the data. In database parlance, this is metadata (data about the data). The lithologic log is geologic information collected during drilling operations. In geologic mapping, lithologic and stratigraphic information is the primary geologic information collected. The lithologic data allows stratigraphic interpretations to be made. Stratigraphic units are the fundamental map unit used in constructing geologic maps.

Metadata includes information about location of drilling, identification of the driller and drill method, date of drilling, site information (elevation, topographic setting, and weather conditions), and geologist of record. Perusal of several hundred records indicates that the metadata of drill holes has varying levels of completeness and, therefore, quality. The more complete the metadata is, the more useful the data record. Ultimately, the single most

essential element of metadata is information about location, i.e. positional information. Because every drill hole is unique, it should be possible to identify every drill hole by its location. The most common expression of location is in latitude and longitude, but there are other common coordinate systems (state plane, UTM). Positional information in many drill logs is severely lacking. In some cases, it may be a verbal description (300 feet past the intersection on the south side of the road). Because the subsurface lithologic composition is extremely heterogeneous, an incorrect location by as little as 100 ft could affect a site-specific issue. Since the 1990's, geologists at SCGS have taken a more disciplined approach to data collection, and the quality of metadata is significantly improved, primarily through the use of standard coordinate systems and GPS instruments to locate drill-hole sites. Therefore, implementing data standards for database development will have the added value of ensuring quality metadata. If there are required fields to enter, then it is incumbent upon the recorder to make sure the header information is accurate and complete before it goes into the database. A more complete header fulfills two functions. First it provides a more accurate and robust description of the drill site. Second it allows the relational aspect of the database to be employed by giving you more criteria to choose from during queries.

Geologic data collected by geologists, for the most part, is complete and of high value. Most drill logs are developed by registered professional geologists, and SCGS Coastal Plain geologists have vast experience in drilling studies over many years and mapping projects. The only short coming or critique one might offer concerns the consistency of data recording. Different geologists tend to emphasize different attributes while logging a drill hole. The ideal drill log should contain all the properties noted by all the various geologists. Again, the

development of data collection standards would increase the consistency and hence the value of data recorded at a site. Why is this so important? Once you drill a hole, efficiency (doing what's right — make a log), effectiveness (doing the right thing – make a complete log), and economics (doing it at the least cost – do it only once) demand that you extract the maximum amount of information, because once it is done, there may never be a chance to re-drill that same hole. This is one of the reasons SCGS maintains a core repository for which we are also interested developing a database. The core repository provides a way to recreate the original drill hole and provide a means for more in-depth studies (chemical analyses of samples), and it is an inexpensive way to conduct preliminary studies to assess the potential for future investment.

Data standards allow data quality to be evaluated. Because of the intrinsic value of any data collected, one option not on the table is to discard the data. It's all relevant, but, if its quality is deemed to be low, it can be tagged as such in the database. Then it is up to the end user to assess the reliability and usefulness of such a record. For instance, if the location of a drill hole cannot be accurately located because the location description is verbal, then an accuracy level can be assigned to its position. If the location of a drill hole is determined using GPS, the accuracy of this location can receive a higher rating. This system also identifies problematic data more quickly and systematically, which could lead to the rehabilitation of bad data. In some cases, it may be possible to recreate original metadata records by cross referencing them with original field notes and discussions with other geologists, drillers, or agencies.

Implementation

The primary result of this project is a completed data dictionary for SCGS drill records. The next step in developing a drill-hole database is to submit the data dictionary to the DNR's Technology Development section (TD). The dictionary and additional input, such as a preliminary data model, will provide database programmers with the information needed to design a database.

Once the database design is finished, the next phase is data entry. Prior to any data entry, SCGS must prioritize its data needs. Project priorities in the next few years will be along the lower Savannah River. The development of digital data in this area will have a high priority. The data will be used to generate 3-dimensional, subsurface stratigraphic models of the Coastal Plain, and these models will assist in future ground-water studies.

The data dictionary and its development lead to another positive outcome. A fuller understanding of the data leads to developing data standards. In essence, data standards are a manual that documents your procedures for data collection. By formalizing data collection methods, it opens your methods to review by users. Formalized collection instills a discipline and order to collecting data, thus ensuring quality and consistency. It demonstrates to other people how your work is done, and it bestows an imprimatur of legitimacy to your methods.

This leaves the final goal of generalizing the above processes to develop a data dictionary for another group within DNR. So far, parts of this process have been described and analyzed, but a coherent step-by-step process has yet to be established. The difficulty is that

generalizing the process may result in an overly simplistic model. To evaluate the general data-dictionary design, a literature search was made to explore what other groups have done. Search results show that many organizations are quite advanced when it comes to describing their data. Many groups have established working committees to develop data dictionaries, formalize their data gathering procedures, and write guidelines for data dictionaries. One example is the American Health Information Management Association (AHIMA). The development of electronic databases in the medical professional is currently a national issue, and AHIMA has developed a sensible approach to developing a data dictionary (Appendix 3). Therefore, rather than creating a completely new process, AHIMA's process has been adapted, and the modified process is presented in Appendix 4.

The general process developed by this project is divided into three parts, which are succinctly described as input, process, and output. Input, or information intake, consists of gathering information for the data dictionary. Input is collected through searching for data sets, reviewing data records, and interviewing data collectors. Processing involves analyzing the input, recognizing data fields (see Figure 1), and describing and defining those data fields. The output is the final assembly of the data dictionary and initiating data standards.

Potential costs to execute this project are addressed in general terms. Database programming costs cannot be evaluated by this project. There is a potential cost savings of developing the data dictionary internally because it saves the programmer one step. A more detailed analysis could look at this issue to evaluate those savings, as well as the issues of potential time savings and efficient use of DNR resources. A basic cost analysis is presented for

populating the drill-hole database with data. The following analysis provides some idea about the time and cost to fill the database (Table 1).

6000 records		Approximate
25 records/day		
40 days/1000 records		
160-200 days=1500 hours		
$9/hr		Student labor
$13,500		
$2,430		18% benefits
	$15,930	**TOTAL**

Table 1 Cost estimate for data entry

There are two potential obstacles to the final development of the drill-hole database, and they both involve database design and programming. First, the data dictionary may be incomplete, so that the programmers require further clarification and enumeration. This would be fairly easy to resolve because it would only require follow up between the data-dictionary author and the database programmers. All interested parties are available for consultation, and this first problem seems manageable. The second problem is potentially much greater, and it concerns TD (Technology Development) not being able to proceed with developing the database due to manpower, money, or other priorities. To this end, discussions with TD-Program Director and LWC Acting Deputy Director have been initiated. Both have indicated their full support to the development of a drill-hole database and that it will have a high priority (Appendix 5). Alternatively, a way to avoid tying up TD's resources would be to seek outside funding for the programming. As natural resource and energy issues come to the forefront in the state and nation, this could be possible. For the Geological Survey, the 2005 Energy Act has

several provisions in that will eventually provide funds to state surveys to develop databases. How soon those funds become available is unknown at this time.

Project Evaluation

The results of this project can be evaluated and measured quite simply by whether the project goals are reached. By this measure, the completion of the drill-hole data dictionary is a preliminary success. The generalization of the data-dictionary process can only be partly evaluated at this point. The process has been developed, but now it needs to be implemented before a result can be evaluated. The development of the end product, a drill-hole database, is an even longer term measurement, and the total success of this project should only be assessed after these two longer term phases are completed or attempted. Therefore, the success will be best evaluated by examining a series of benchmarks over a period of time (Appendix 6).

Summary and Recommendations

Summary The results of this project provide one firm conclusion. Individuals can take control of their data and begin the process of database development on their own. There is something empowering about this result. It shows that data originators are able to take a leadership role in the database development of their own data.

Recommendation A generalized process for developing a data dictionary has been proposed, it needs to be tested with a data set outside the Geological Survey. Additionally,

other data sets at the Geological Survey will also be dictionized and prepared for database development (e.g. map-image library, core repository, publications.

Recommendation Development of data standards should take a high priority. Because the data dictionary process can lead to a greater understanding of your data, it would be advantageous to document those procedures and standards. Using standards are a front-line quality assurance. When combined with data entry through a database, the result is an active QA/QC program.

Recommendation Integrate drill-hole database with 3-Dimensional mapping software. Water availability issues are moving to the top of many Agency, county, and municipal problem lists. Using 3D maps will provide an accurate delineation of the geology and structure of an area, which, in turn, provides a better understanding of the subsurface hydrologic system.

Recommendation With the possibility of more databases within the Geological Survey, it might be prudent to investigate hiring a short-term database administrator. This person would be dedicated to working the rough edges off the Survey's databases, provide routine maintenance, and establish connections with other DNR databases. Once the Survey's databases are considered fully operational, the database administration could be switched to TD. At this time, the only way to fund such a position is with outside funds. It will be necessary to find a funding source.

GLOSSARY

Coastal Plain - A low, generally broad plain that has its margin on an oceanic shore and its strata either horizontal or very gently sloping toward the water, and that generally represents a strip of recently prograded or emerged sea floor, e.g. the coastal plain of SE U.S. extending 3000 km from New Jersey to Texas. Its inland limit is the Fall Line where large streams cascade off the more resistant metamorphic rocks of the Piedmont (e.g. Columbia, Washington, DC, Richmond). (AGI Glossary of Geology)

Data dictionary - "A data dictionary is a collection of descriptions of the data objects or items in a data model for the benefit of programmers and others who need to refer to them. A first step in analyzing a system of objects with which users interact is to identify each object and its relationship to other objects. This process is called data modeling and results in a picture of object relationships. After each data object or item is given a descriptive name, its relationship is described (or it becomes part of some structure that implicitly describes relationship), the type of data (such as text or image or binary value) is described, possible predefined values are listed, and a brief textual description is provided. This collection can be organized for reference into a book called a data dictionary."
http://searchsoa.techtarget.com/sDefinition/0,,sid26_gci211896,00.html (last accessed 1/28/08)

Data element - A single component of information from a sample location. It identifies a single attribute about the sample site. See Figure 1

Data field- A field is an area in a fixed or known location in a unit of data such as a record, message header, or computer instruction that has a purpose and usually a fixed size. In some contexts, a field can be subdivided into smaller fields. See Figure 1 http://searchoracle.techtarget.com/sDefinition/0,,sid41_gci213963,00.html (last accessed 1/28/08)

Data hierarchy - An arrangement of data consisting of sets and subsets such that every subset of a set is of lower rank than the set. See Figure 1

Data model- A description of the organization of a database. It is often created as an entity relationship diagram. Today's modeling tools allow the attributes and tables (fields and records) to be graphically created. http://www.pcmag.com/encyclopedia_term/0,2542,t=data+model&i=40815,00.asp (last accessed 1/28/08)

Data record - A collection of information about a specific site. It is a collection of data elements that describes the information collected at a site for a project. See Figure 1

Data table - A group of one or more data records. See Figure 1

Drill Log or Lithologic Log - A continuous record as a function of depth, usually graphic and plotted to scale, of observations made on the rocks or sediment (lithologic log) of the geologic section exposed in a drill boring. (AGI Glossary of Geology)

GIS - "A GIS (geographic information system) enables you to envision the geographic aspects of a body of data. Basically, it lets you query or analyze a database and receive the results in the form of some kind of map. Since many kinds of data have important geographic aspects, a GIS

can have many uses: weather forecasting, sales analysis, population forecasting, and land use planning, to name a few."

http://searchsqlserver.techtarget.com/sDefinition/0,,sid87_gci213982,00.html (last accessed 1/28/08)

GPS - "The GPS (Global Positioning System) is a "constellation" of 24 well-spaced satellites that orbit the Earth and make it possible for people with ground receivers to pinpoint their geographic location. The location accuracy is anywhere from 100 to 10 meters for most equipment. Accuracy can be pinpointed to within one (1) meter with special military-approved equipment. GPS equipment is widely used in science and has now become sufficiently low-cost so that almost anyone can own a GPS receiver."

http://searchmobilecomputing.techtarget.com/sDefinition/0,,sid40_gci213986,00.html (last accessed 1/28/08)

Lithology – The description of rocks, esp. in hand specimen and in outcrop, on the basis of such characteristics as color, mineralogic composition, and grain size. (AGI Glossary of Geology)

Metadata - "Meta is a prefix that in most information technology usages means "an underlying definition or description." Thus, metadata is a definition or description of data and metalanguage is a definition or description of language."

http://searchsqlserver.techtarget.com/sDefinition/0,,sid87_gci212555,00.html (last accessed 1/28/08)

QA/QC - "Quality control (QC) is a procedure or set of procedures intended to ensure that a manufactured product or performed service adheres to a defined set of quality criteria or

meets the requirements of the client or customer. QC is similar to, but not identical with, quality assurance (QA). QA is defined as a procedure or set of procedures intended to ensure that a product or service under development (before work is complete, as opposed to afterwards) meets specified requirements. QA is sometimes expressed together with QC as a single expression, quality assurance and control (QA/QC)."
http://whatis.techtarget.com/definition/0,,sid9_gci1127382,00.html (last accessed 1/28/08)

Relational Database - "A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables."
http://searchsqlserver.techtarget.com/sDefinition/0,,sid87_gci212885,00.html (last accessed 1/28/08)

Stratigraphy - The science of rock strata. It is concerned not only with the original succession and age relations of rock strata but also with their form, distribution, lithologic composition, fossil content, geophysical and geochemical properties – indeed, with all characters and attributes of rocks as strata; and their interpretation in terms of environment or mode of origin, and geologic history. (AGI Glossary of Geology)

REFERENCES

Bates, R.L., and Jackson, J.A., eds., 1987, Glossary of geology (3rd edition): Alexandria, Va., American Geological Institute, 788p.

AHIMA e-HIM Work Group on EHR Data Content, 2006, Guidelines for Developing a Data Dictionary: Journal of American Health Information Management Association, v. 77, no.2 (February 2006), p. 64A-D.

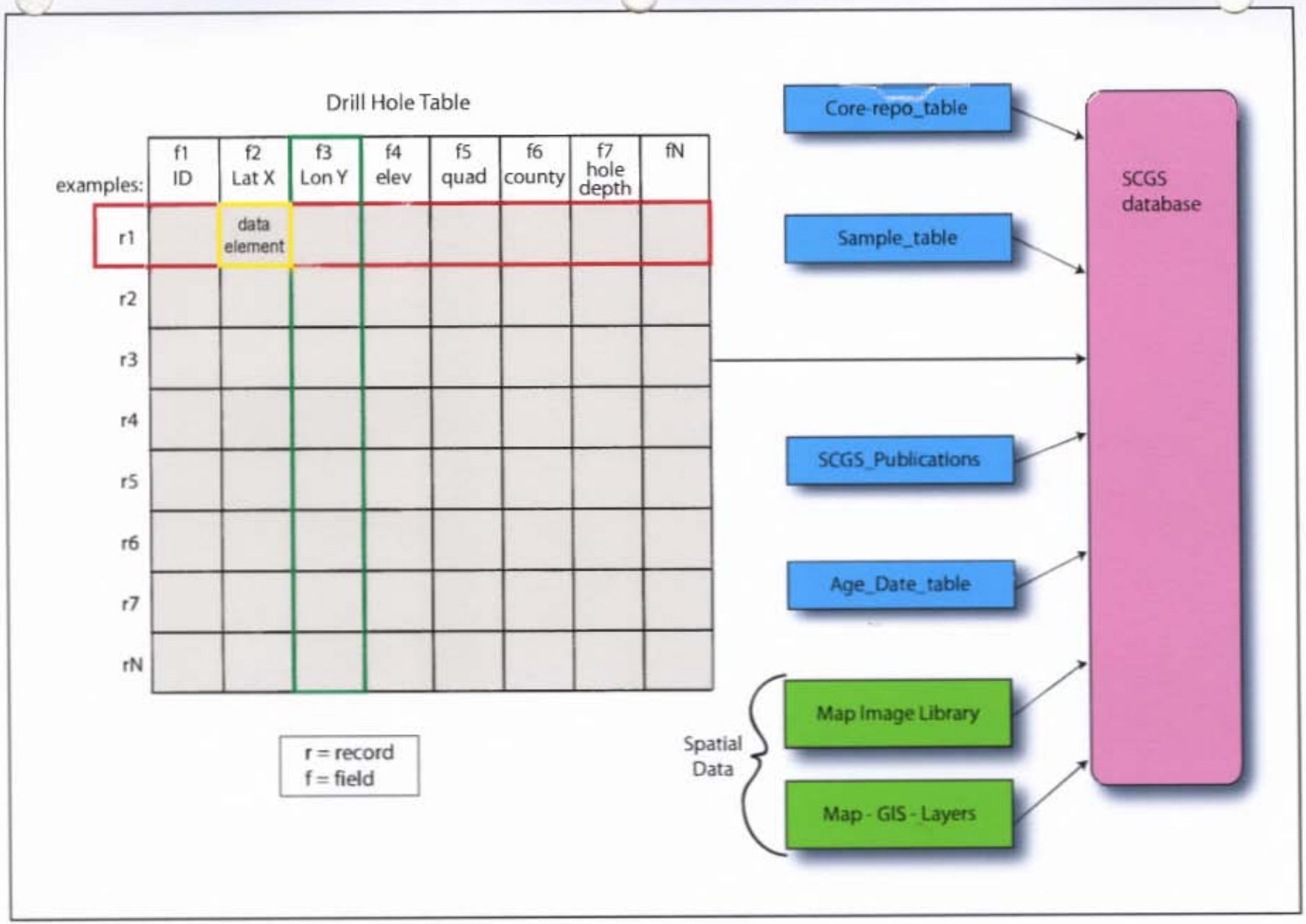


Figure 1. Schematic diagram of data sets at SCGS showing preliminary data hierarchy of drill-hole table.

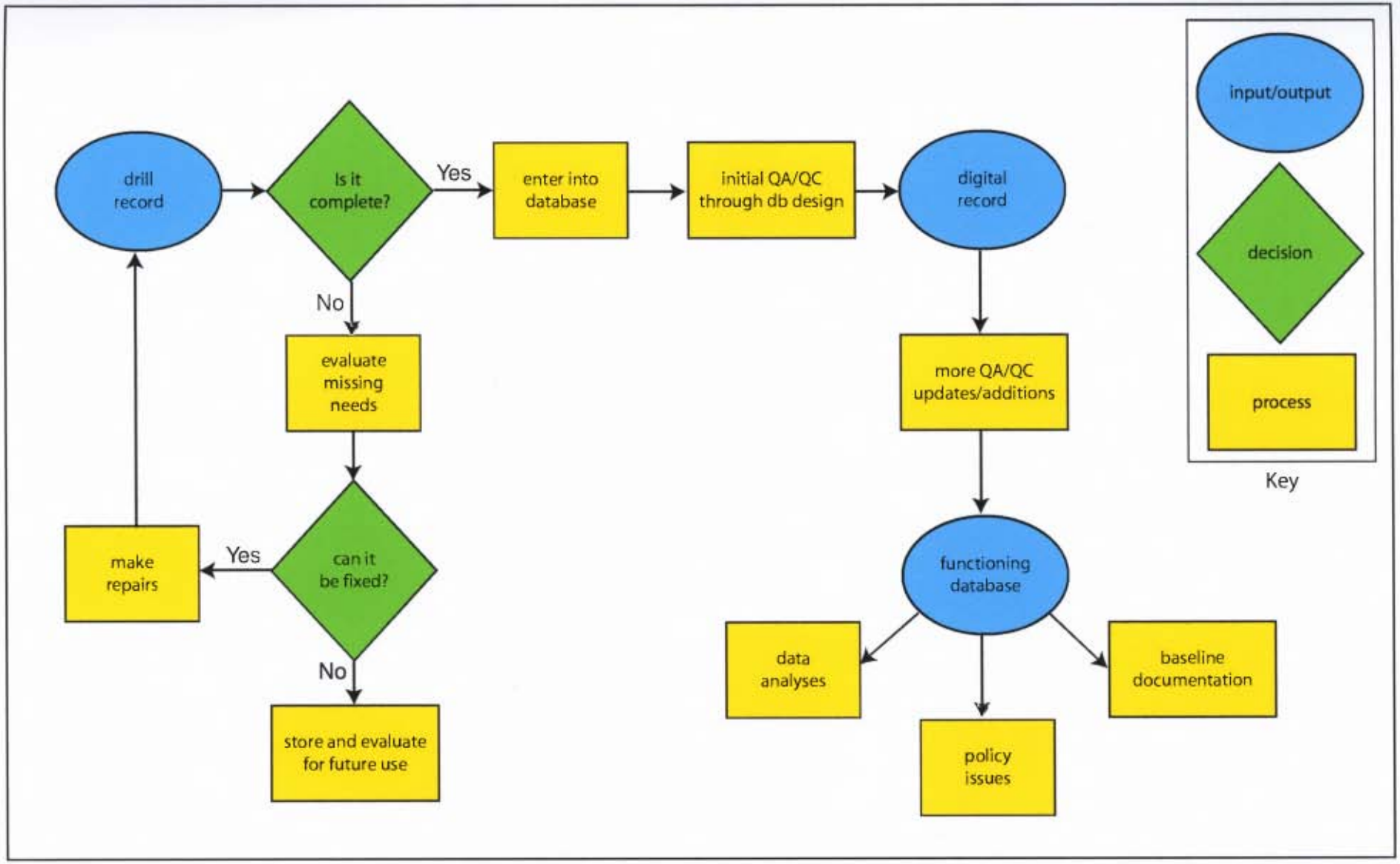


Figure 2. Flow Chart of SCGS drill-hole record entry into database

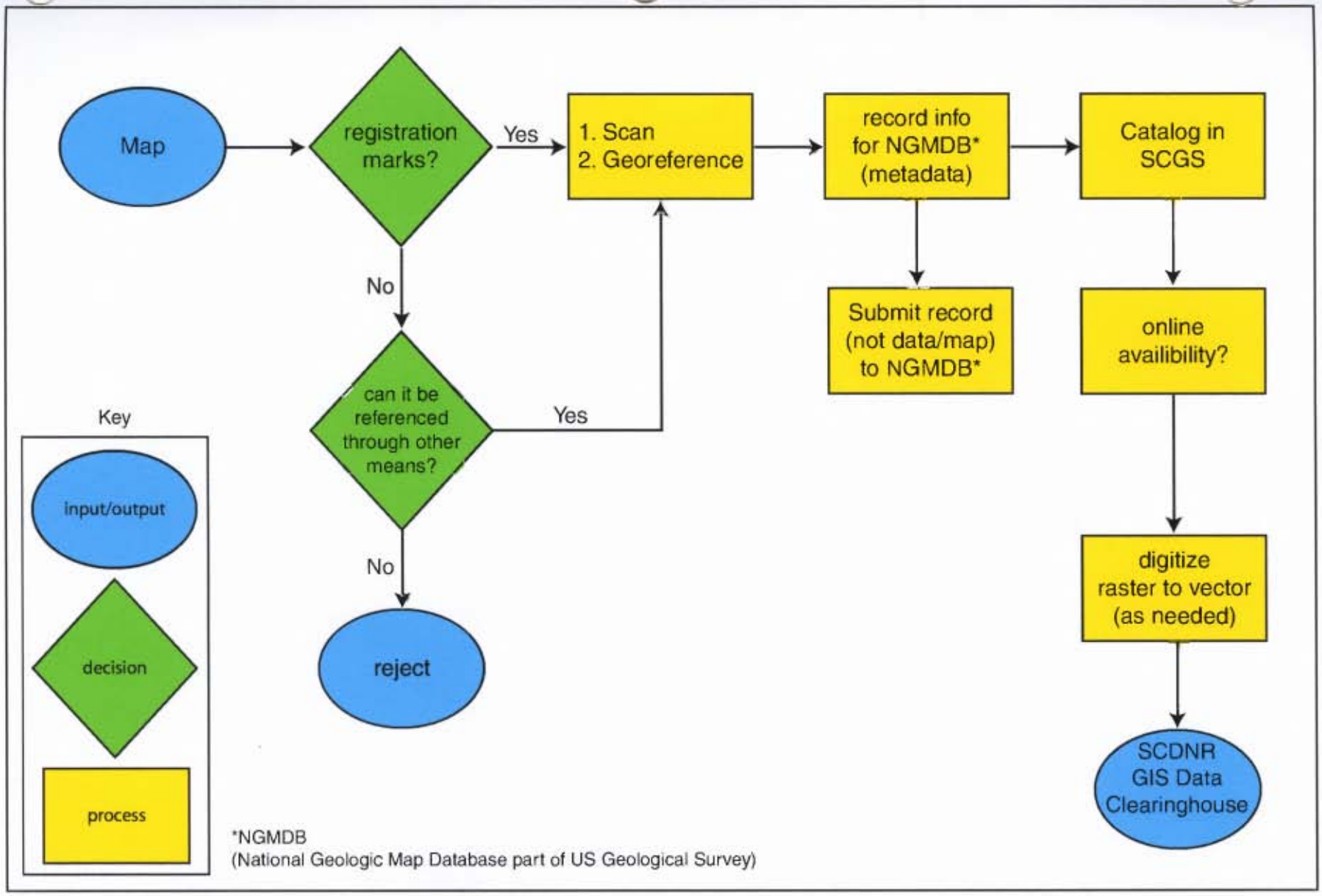


Figure 3. Flow Chart of Map Image entry into database

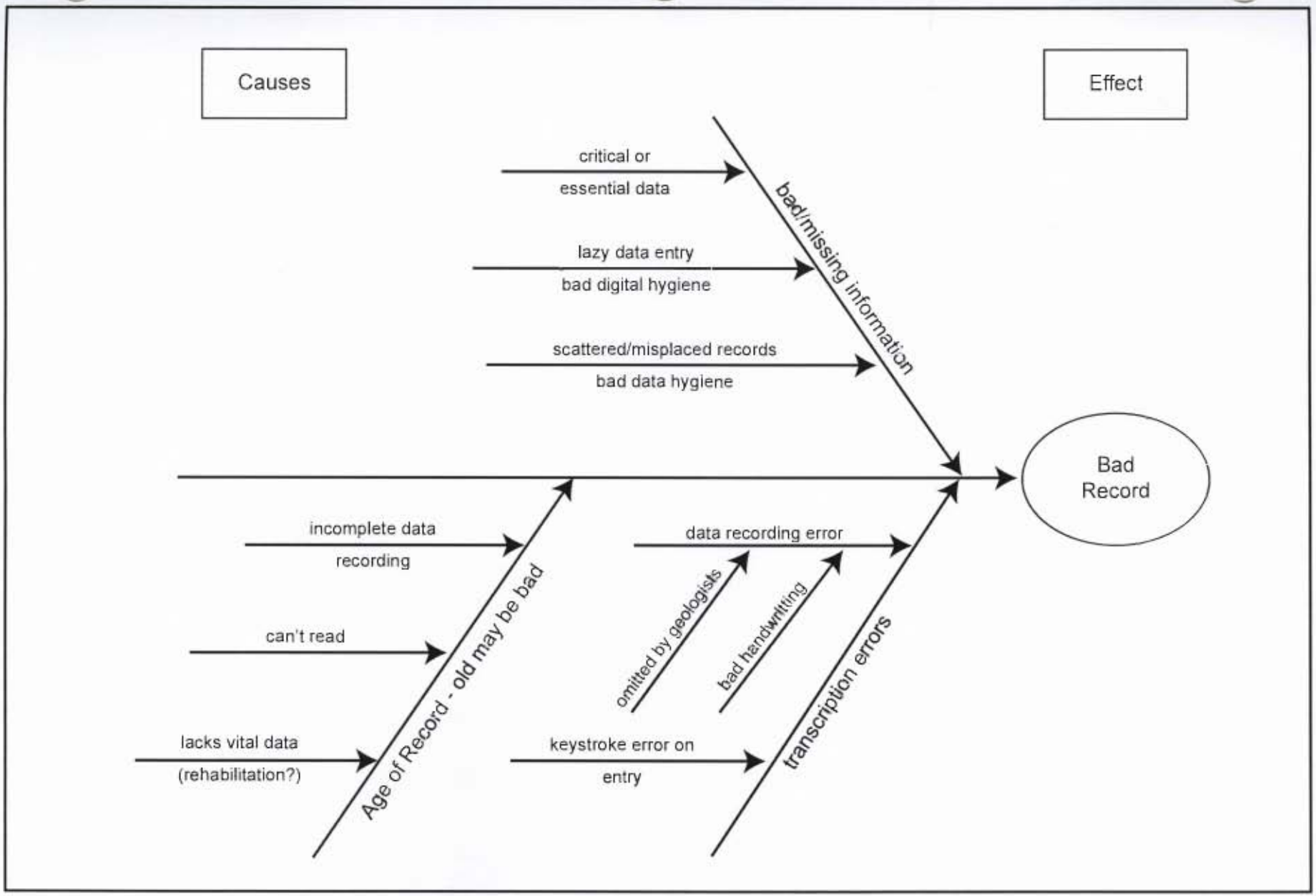


Figure 4. Cause-effect diagram tracing possible error sources in a bad data record

Appendix 1

Sample Lithologic Log from SCGS Drill-hole Library

Auger Log -- 40-175			
SCGS ID#:	40-175	**TD:**	60 ft / 18.3 m
7.5' Quad:	Wateree	**Collar:**	161 ft / 49.1 m
County:	Richland	**Date:**	01/06/05
Field ID:	W9	**Type:**	Power Auger
Coordinates:	3745612 N / 528487 E	**System:**	UTM Z17N, NAD27
Location:	North shoulder of Webbrook Road near the intersection with Poultry Lane; across from vacant lot just east of #236		
Drilled by:	Gary Taylor (driller), Ernest Howard Jr., Quinton Jones		
Logged by:	David C. Shelley		
Remarks:			

Summary of Units

Unit	Thickness (ft/m)		Depth (ft/m)		Elevation (ft/m)	
Congaree River Valley Terrace	40	12.2	0-40	0-12.2	121-161	36.9-49.1
Kuu	20	6.1	40-60	12.2-18.3	101-121	30.8-36.9

Log

0'-5' (156' -161'): brown (10YR4/6) grading down to brownish orange (7.5YR6/8) @ 2-5', poorly sorted, sub-ang to sub-rnd, f.s. to m.s.; com. interst. clay, mod. cohesive, crumbly; grades down to mod. abnt. interst. clay

5'-11' (150' -156'): brown (10YR4/6), brownish orange (7.5YR6/6), and light grayish brown (2.5Y7/2) interlaminated, grading to all orange brown (10YR5/6)@ 10-11', mod. well sorted, sub-rnd to well round; no hm, rare micas; rare interst. clay, mod. cohesive, crumbles

11'-20' (141' -150'): brownish orange (7.5YR6/6), with brick red (10R5/8) and white (2.5Y8/1) streaks; clay; gritty; micaceous; some very fine hm

20'-22' (139' -141'): brownish orange (7.5YR6/6), mod. well sorted, sub-ang to sub-rnd, f.s.; occasional c.s. to v.c.s.; qtz, com. hfm, unc. micas; unc. interst. clay, sl. cohesive, plastic, clumpy

22'-25' (136' -139'): brownish orange (7.5YR6/6), mod. poorly sorted, sub-ang to sub-rnd, f.s. to c.s.; qtz, unc. hfm; very com. interst. clay, sl. cohesive, damp

25'-40' (121' -136'): pale brownish orange (10YR7/6) grading down to brownish orange (7.5YR7/6) @ 30-36, grading down to white (N0) @ 36-40, mod. sorted, sub-ang to sub-rnd, m.s. to v.c.s.; qtz, unc. hfm; rare interst. clay, sl. cohesive, wet, mushy; @ 30-36', wet and soft, with c.s. to v.c.s. wit h unc. interst. clay and rare grav. <1.0 cm; @ 37-39', lag, with blocky grav. <4.0 cm; some thin clay stringers @ 37 -39', but these may be contamination

40'-42' (119' -121'): dark brownish orange (7.5YR5/8), mod. sorted, sub-ang to sub-rnd, m.s. to c.s.; qtz; ubiq interst. clay, very very cohesive, plastic, wet

42'-53' (108' -119'): orange (7.5YR7/8) grading down to brownish orange (10YR6/8); clay; gritty; micaceous; sparse, sub-ang to sub-rnd meduim to c.s.; very cohesive; variably damp and softly plastic to dry and stiff; grades to clay matrix supported, poorly sorted sub-ang to sub-rnd f.s. to coarse qtz sand @ 52-53'

53'-54' (107' -108'): brownish orange (10YR6/8) with pale yellow (2.5Y8/4) and brick red (10R5/6) highlights, mod. well sorted, sub-ang to sub-rnd, f.s. to m.s.; qtz, com. hm, unc. micas; sl. cohesive, crumbles; some thin clay stringers

54'-60' (101' -107'): white (N0), mod. well sorted, m.s. to c.s.; rare v.c.s.; qtz, com. hm, com. hfm; abnt. interst. clay, poorly cohesive, mushy; forms gray patina on rods

Appendix 2

Data Dictionary of SCGS Drill-hole Library

Name	Type	Description	Full Name
recordby	text	who entered this data, need an index of current and past data enterers, Initials	record by
date_filed	date	date digital record was entered, also would be last date of modificiation (automatic?); MM, DD, YYYY; if month or date unknown, use 00; year must be entered	date entered
datasource	text	source of data, usually SCGS, but SCDNR-hydro (SCWRC), DHEC, USGS, owner, other, unknown; who submitted this data for entry	source of data
update	Y/N	is this record an update of a preexisting record	update
		Preliminary is a book keeping set of entries. It tells who entered the data into the DB, when it was done, what they used to enter the data, and whether the entry is an update. Information about the data entry is necessary for a few reasons. First, you	

Drill-Hole Data Dictionary — Owner

Name	Type	Description	Full Name
Owner first	text	Site Owner firstname	First name of owner
Owner Last	text	Owner last, could be an Agency or Company, City, State, County entity? Or unknown	Last name or Agency name
Address 1	text		street address
Address 2	text		secondary street address
Address 3	text		secondary street address
City	text	City	City
State	text	State	State
zip	number	fixed field for postal zip code	zip code
		should there be an index of owners?	
		at this time, other than SCGS, who would other owners be?	
		Information about the owner of the drill hole is located here. In most cases the owner is SCGS. Other possiblities could be USGS, local well drillers, or some other governmental agency. Drop down pick lists could be developed for the usual suspects.	

Drill-Hole Data Dictionary — Site

Name	Type	Description	Full Name	Required?
SCGSID	text	Unique SCGS ID, combination of 2 numbers, 1st= county code; 2nd =hole number for county	SCGS Identification Number	Y
LocalID	text	field designation of drill hole, given at time of drilling, could be anything, non-unique from quad to quad or project to project	Local or Field Designation	
DrillLocX	number	X coordinate describing geographic location of core, either LON or UTM_Easting. Because LON will probably have to be parsed, make this a text string??	X coordinate (easting or LON)	
DrillLocY	number	Y coordinate describing geographic location of core, either LAT or UTM_northing. Because LAT will probably have to be parsed, make this a text string??	Y coordinate (northing or LAT)	
Datum	text	horizontal datum of geographic coordinates; NAD 1927, 1983, others?	horizontal datum	
CoordMethod	text	How were map coordinates determined? Located by sketch map, located in field on map (scale?), GPS, differential GPS, survey, calculated from other coordiantes, approximate, unknown	Coordinate Method	
CoordAccuracy	text	evaluation of coordinate method; possibly combine as one with method, but two fields may be more flexible in searches.	Coordinate Accuracy	
Quad	code	SC 7.5-minute topographic quadrangle, 5-letter abbreviation code as used by DNR	Quadrangle Name	
County	code	SC 3-letter County abbreviation	County	
RiverBasin	code	Major river basin delineation; major basins only	Major River Basin	
SurfaceElev	number	Surface Elevation of drill site, altitude above sea level, measured in feet	Surface Elevation	
ElevMethod	text	method for deteriing elevation, map, survey, gps, diff gps, altimeter, calculated (interpolated from DEM), unknown	Elevation method	
ElevAccuracy	text	accuracy of elevation method; table of possible vaules. Greatest might be map, which is typically 1/2 contour interval	Elevation Accuracy	
PhysiographicReg	text	Piedmont or Coastal Plain or Triassic?	Physiographic Region	
TopoSetting	text	brief description of topo setting: interfluve, hilltop, dune, flood plain, terrace, sinkhole, depression, stream channel, etc	Topgraphic Setting	
Project	text	Project for which drill hole was made; drill holes are invariably associated with a specific project. Easy way to collect information about a project.	Project	
questionable				
state		To distinguish between SC and neighboring states. Is this necessary?	state	
		Site data is information about the drill site. It is sometimes referred to as header data, as it is found typically at the top of a field lithologic log. It consists of the the two most important data elements: SCGSID and Location XY. Without this inf		
		After talking with Jim S., it is best to pick one coordinate system (LL or UTM) and stick with it. Because a majority of work (all?) is done in UTM and folks are familiar with it, it makes sense to stick with UTM. All data should be entered in UTM.		

Drill-Hole Data Dictionary — Drill Hole

Name	Type	Description	Full Name
DrillerID	text	Drilling Company, Contractor	Driller Identication
DrillMethod	text	Drilling method, auger, hydraulic rotary, diamond coring, dug, jetted	Drilling Method
DrillDate	date	date of drilling, if drilling over several days, the first day of drilling	Date of Drilling
DrillDepth	number	depth to bottom of hole	Depth of Drilling
Diameter	number	diameter of hole	Diameter of borehole
DataType	text	type of drill hole, research, exploration, test well, production, observation, test hole	Data Type
Geologist	text	Geologist of record, need and index of current and past geologists	Geologist
Driller	text	Driller of record	Driller
Project	text	Code designation for project, STATEMAP, SRS, etc.	Project
DrillLog	Y/N	Lithologic log by driller, unaccredited geologist?	Driller Log
GeolLog	Y/N	Lithologic log by geologist	Geologist Log
GeophysLog	Y/N	Geophysical Log type, pull down menu?	Geophysical log
SiteStatus	text	what became of the this hole, in use, abandoned, filled in, not in use, standby, unknown	Site Status
SiteUse	text	Use or purpose of well, eng/test boring, geo/hydro research, observation, oil/gas, test, withdrawal, other	Site Use
WaterUse	text	agriculture, commerical, domestic, fire, geothermal, industrial, observation, other, public supply	Water Use
Samples	text	index: cuttings, grab, split spoon, core (a null field means no sample)	Sample
StatWat	number	static water level measured in open drill hole	Static Water level
		Important information about the drill hole. It gives a quick snap shot about the hole, how big, how deep, when drilled, who did it, who logged it, was there any special work done on the hole, e.g. well installed, geophysical tests. Much of the usefulnes	

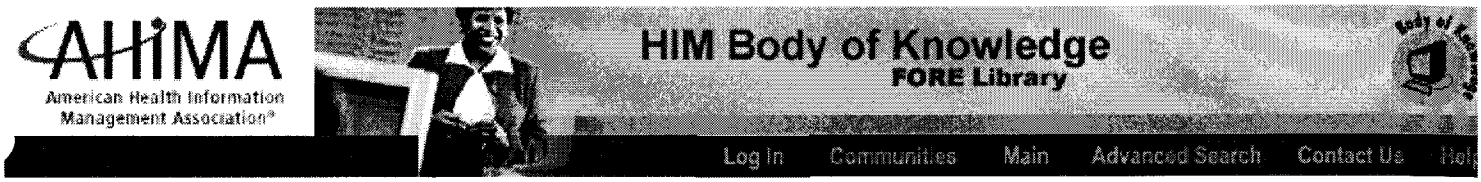
Drill-Hole Data Dictionary — Lithology

Name	Type	Description	Full Name
IntTop	number	Elevation of top of lithologic interval, measured as feet below surface?	Top of Interval
IntBot	number	Elevation of bottom of lithologic interval, measured as feet below surface?	Bottom of interval
Litho	text	Description of geologic material in this interval, possibly develop pull down menu?	lithologic description
text	text	texture of sediment	textural description
sort	text	description of sorting, pull down, well, moderate, poorly	sorting
round	text	description of grain roundness	roundness
color	text	primary color of material	color
lab_field	text	lithologic description based on field or lab analysis	Lab or Field description
GeoFM	text	Intepretation of geologic formation, library of fm abbreviations	Geologic Formation
		might be better to have only one field for text/lith/sort/round/etc and then add a field on whether measurement is from field or lab but what if it's both? don't need to worry about that for the dictionary, it's the database that will sort that out	
		Still haven't sorted out this table yet. The description of material may be easier by itself, then add the tag that it's a field or a lab description. This would keep repetitions and empty fields to a minimum.	

Appendix 3

Guidelines for Developing a Data Dictionary

Developed by AHIMA



Guidelines for Developing a Data Dictionary (AHIMA Practice Brief)

Information systems are only as good as their data. Without a mutually agreed-upon set of data elements with clearly defined names and definitions, the validity and reliability of the data contained in a system are suspect at best and must be discounted at worst. The data dictionary and its relationship with the metadata registry are the foundation of an information system and the central building block that supports communication across business processes.

Data Dictionary Defined

To advance work toward electronic health record (EHR) content, AHIMA formed an e-HIM® work group to educate members and the industry on the importance of standardizing data content and data definitions within provider organizations and the industry as a driver to quality of care and patient safety. The work group defined a data dictionary as a descriptive list of names (also called representations or displays), definitions, and attributes of data elements to be collected in an information system or database. The purpose of the data dictionary is to standardize definitions and ensure consistency of use.

Rationale for Data Dictionaries

Standardizing data enhances interoperability across systems. It also improves data validity and reliability within, across, and outside the enterprise. Communication is improved in clinical treatment, research, and business processes through a common understanding of terms. Standardization provides developers with a common road map to promote consistency across applications.

Lack of a sound data dictionary can cause problems within and across organizations. Organizations may call the same data element by different names or they may call different data elements by the same name across an enterprise. As a result, an organization may not collect all of the information it needs or it may be unable to combine or map data across systems because the definitions are not identical. A worse possibility is that an organization may combine data elements it believes to be equivalent and draws incorrect inferences from the invalid data. Multiple users entering data may have different definitions or perceptions of what goes into a data field, thereby confounding the data (e.g., are "reason for visit" and "chief complaint" the same or different?).

Large complex systems with multiple stakeholders (internal and external) often require use of multiple, differing data sets. Variances among the data sets that are not recognized across the system can affect the information flow as well as the workflow. Maintaining expansive, overlapping data sets is costly to the organization in time and money and affects the quality of care. The organization will not be positioned for harmonizing information at the regional or national level.

Guideline Development Process

The work group conducted a comparative study of data definitions at the field definition level in order to create guidelines for developing a data dictionary. The purpose of these guidelines is to assist in building data dictionaries at the organizational level, aid in the development of new and existing data content standards, and support national standards harmonization efforts.

Since it is too early to know the impact of the federal data standards harmonization project sponsored by the Office of the National Coordinator for Health Information Technology, the work group centered its work around data dictionary development, whether new or updated. It is not too early for organizations to clean up their own houses through alignment of data content; this optimizes internal understanding as well as prepares for further alignment with the federal effort.

Taking care to select data sets affecting all care settings, the work group chose the following 11 major industry standard data sets for comparison:

- ASTM International's E1384-02a Practice for Content and Structure of the Electronic Health Record Minimum Essential Data Set
- ASTM International's WK4363 Standard Specification for the Continuity of Care Record (CCR)
- Doctor's Office Quality Information Technology's Data Element Specification v.1.1.2
- Electronic Medical Summary project (Canada) Core Data Set
- International Organization for Standardization (ISO)/TS 18308 Health Informatics: Requirements for an Electronic Health Record Architecture
- Joint Commission on Accreditation of Healthcare Organizations Comprehensive Accreditation's Manual for Ambulatory Care: Information Management Standards 6.20, EP1
- Centers for Medicare and Medicaid Services' Minimum Data Set, Version 2.0, for Nursing Home Resident Assessment and Care Screening
- National Center for Vital and Health Statistics' Core Health Data Elements
- Centers for Medicare and Medicaid Services and the Joint Commission on National Hospital Quality Measures
- AHIMA's Personal Health Record Minimum Common Data Elements
- Health Level Seven's Clinical Document Architecture, release 2

The work group selected common data content standards to compare at the field level. The group agreed that a sample of 10 data elements would be selected from each of a variety of data category types (e.g., service instance, patient, observation, providers, orders, care, treatment plan, encounter, problems) for comparison across the selected data sets. Initially, the work group chose ASTM International's CCR as a base data set from which to select a representative sample of data elements. It quickly became apparent that ASTM International's E1384-02a Minimum Essential Data Set was far more developed in detail and inclusiveness. As a result, it became the base against which other data sets were compared.

Using the information gained from this comparative study along with their collective expertise, the work group created the following guidelines to assist the industry in the development of data dictionaries.

Guidelines for Developing a Data Dictionary

1. Design a plan for the development, implementation, and continuing maintenance of the data dictionary.

Preplanning is imperative. The development of a data dictionary is part of a larger process. An information model must first be developed to align the workflow with information flow. This includes deciding what data are required, how the data will be used, who will use the data, and how the data will flow internally and externally, including communications with other entities.

This should be a collaborative process, and stakeholders should be encouraged to resist the temptation to collect data simply because they can. In the ideal scenario, data are captured once for use by multiple users. The end result of this data mapping is the ability of multiple entities to mine the same data source. Each will know the exact nature of the data element each is accessing. The plan should also include:

- The type of media (paper, electronic, spreadsheet, relational database) in which the data dictionary

will be developed and maintained. The media choice may depend on the complexity of the enterprise system and the availability of resources.
- Adequate funding and staffing with clearly defined roles and responsibilities for development and ongoing maintenance of the data dictionary. Databases are dynamic and can be affected by new business lines, changes in national standards, and clinical advancements.
- Provisions to ensure that all licensing agreements are in order.
- Ongoing education and training of all staff as appropriate to their use of data elements and their definitions.

2. Develop an enterprise data dictionary that integrates common data elements used across an enterprise.

One purpose of the data dictionary is to provide consistency and understanding of common data across applications. Preplanning is a must to accomplish this at an enterprise level. A process must be clearly defined and key stakeholders identified. The process requires collecting information or metadata (data about the data) on each data element found to be common across domains. It is important to define up front what needs to be done before starting the dictionary. This includes defining what metadata will be collected on each element as well as what will not be collected. Examples of metadata include name of element, definition, application in which the data element is found, locator key, ownership, entity relationships, date first entered system, date element terminated from system, and system of origin.

A metadata registry is an authoritative source of reference information about the representation, meaning, and format of data collected and managed by an enterprise. It does not contain the data itself but the information that is necessary to clearly describe, inventory, analyze, and classify data.

3. Ensure collaborative involvement and buy-in of all key stakeholders when data requirements are being defined for an information system.

Stakeholders include data creators, data owners, and data users, both internal and external to the organization. Representation should reflect all geographies (departments, facilities, satellites, corporate representatives, and external entities). Each organization must identify its stakeholders based on its own unique business model, organizational structure, information flow, and reporting requirements. Different stakeholders may have different data element definitions within their local domain. Every attempt should be made to promote collaborative agreement whereby a datum is collected only once even though it may be used by multiple end users.

Take for example a large enterprise that discovered it had approximately 40 different representations for data elements with a set of values of "yes" and "no" throughout its data dictionary. These included: Y = yes, N = no; yes, no; 1 = yes, 0 = no; 1 = no, 0 = yes; 1 = yes, 2 = no. These should be standardized as one set of values in the enterprise data dictionary.

Public health and research are examples of external stakeholders. Public health reporting is often forgotten in the data requirements definition phase. As a result, organizations incur extra costs to develop special interfaces and maintain crosswalk tables to meet public health requirements.

The collaboration of all data stakeholders (e.g., clinical specialties, support services, HIM services, IS services, reimbursement specialists, administrative, legal, and public health agencies) should enhance consensus and understanding of data and their flow across all domains.

4. Develop an approvals process and documentation trail for all initial data dictionary decisions and for ongoing updates and maintenance.

It is important to document decisions made about the data dictionary throughout the life of the system. Each

subsystem (e.g., finance, lab, radiology) should have one authoritative owner responsible for tracking all implemented data dictionary activations, deactivations, relevant dates, events, and decisions.

There must be a maintenance and change control process for adding new values, elements, and enactment dates. The subsystem owner should review and approve any additions to the system and integrate those changes through a collaborative process with other owners into the whole enterprise system. The process should address how a new datum applies in the local setting or domain and across all aspects of the enterprise.

5. Identify and retain details of data versions across all applications and databases.

Ensure clear mapping instructions for organization-specific definitions. Version control is essential for maintaining data reliability. It is important that the data set version is clearly identified. Differences between versions may be minor or extensive. It is critical that everyone in the enterprise operate on the same version in order to maintain data integrity and continuity. Version control is essential for data dissemination in standard format to satellite or remote facilities. Separate tables may be considered for keeping track of changes such as additions, deletions, and their relative effective dates.

6. Design flexibility and growth capabilities into the data dictionary so that it will accommodate architecture changes resulting from clinical or technical advances or regulatory changes.

Build expansion capabilities into the fundamental design to accommodate a dynamic system. There should be a plan for future expansion, such as expanding a data field from one element to multiple elements. Expansion must be carefully addressed because of the potential ramifications of concept migration, the change of an idea or concept over time through growth or change to the system. This becomes problematic when comparing data across time if the meaning of a particular element has changed while its name or representation has not. If a data element is totally revamped, document when that specific data element went into effect and when it was deactivated. If the data element expands into something new, do not migrate the old concept but rather create a new element to move forward. This will affect how the data are stored and retrieved. It may require consultation with vendors where current system limitations exist.

Always strive for concept permanence. Never reuse a concept even if it becomes obsolete. For example, when an ICD code number is retired, never reassign the retired code to a new concept. Always follow the defined coding practices. This becomes particularly important in data comparison. Address architecture flexibility in vendor contracts to allow for system upgrades and room for expansion to accommodate requirements common to provider-specific issues, user groups (multiple sites), or state-based directives.

7. Design room for expansion of field values over time.

Consider future needs to collapse and expand values to accommodate mapping from a larger to smaller or smaller to larger number of values within a field definition. When setting up the information system, consider how to accommodate multiple systems and how to go from one code system to another. Mapping and transferring guidelines should be clarified between data sets. For example, race or ethnicity is frequently defined with different values. One data set has four items, another has six. The mainframe or core system needs the maximum amount of values. The mapper needs to know the rules to use when collapsing six values into four. Migrating four to six is usually impossible, which creates other issues.

Gender is another core data element that can generate much discussion. Many systems only allow for male and female, while others provide for unknown and other. When an "other" category is an option, there should be a process for monitoring what is captured under that heading. When large numbers begin to appear in the category, there should be a review to determine if a new discrete category is required or if there is misunderstanding in the definition of the core element.

Take for example a data dictionary that must accommodate the changes necessary to adopt the current ICD-9-CM diagnosis code fields from six characters to what will be required for ICD-10-CM. Some organizations have been proactive and already made these changes and updated their data dictionary.

8. Follow established ISO/International Electrotechnical Commission (IEC) 11179 guidelines or rules for metadata registry (data dictionary) construction to promote interoperability and automated data sharing.

Uniformity of approach in data dictionary development avoids industry fragmentation. In an effort to promote and improve international communications among governments, businesses, and scientific communities, ISO and IEC have developed standards for specification and standardization of data elements. The ISO/IEC 11179 standard consists of:

- A framework for the generation and standardization of data elements
- A classification of concepts for the identification of domains
- Basic attributes of data elements
- Rules and guidelines for the formulation of data definitions
- Naming and identification principles for data elements
- Registration of data elements

This standard provides excellent detailed information and examples of how to classify and define data elements. It also includes examples of pitfalls and practices to avoid.

9. Adopt nationally recognized standards and normalize field definitions across data sets to accommodate multiple end user needs.

It is important to define all data characteristics to be included for each data element for all domains. This includes specifying domain boundaries and identifying linkages across domains. This will require extensive discussion and agreement among all stakeholders. The ideal is the development of a common integrated data and terminology model. Terminologies should be coordinated to eliminate overlaps, redundancies, and inconsistencies. This will eliminate the need for mapping among terminologies.

10. Beware of differing standards for the same clinical or business concepts.

Do not assume that things labeled the same are actually identical or will map one to one. For example, there are several different wound staging protocols. The Centers for Medicare and Medicaid Services require one version in the Minimum Data Set Version 2.0 for reimbursement purposes. For clinical care, it requires a different staging protocol that is based on the AHRQ Clinical Practice Guideline for Pressure Ulcers. MDS 3.0, currently in beta with an expected release date in 2007, is expected to remedy this particular problem by requiring only one standard. Pain measurement scales are another example of multiple scales for the same concept. Always check with a subject matter expert to ensure valid data.

11. Use geographic codes and geocoding standards that conform to those established by the National Spatial Data Infrastructure and the Federal Geographic Data Committee, following the guidelines of the Federal Information Processing Standards.

Valid street addresses, zip codes, county, state, and country codes are important to information exchange across systems and geopolitical boundaries. Standardization of geographic codes enhances interoperability of systems. Healthcare uses this information for tracking diseases as well as people. Using internationally accepted standards further enhances the interoperability of systems and the exchange of information. The following are recommended resources for geographic codes:

- Federal Information Processing Standards (www.itl.nist.gov/fipspubs)

- Federal Geographic Data Committee (www.fgdc.gov)
- United States Postal Service (www.usps.gov)
- National Spatial Data Infrastructure (www.fgdc.gov/nsdi/nsdi.html)
- International Organization for Standardization (www.iso.org)

12. Test the information system to demonstrate conformance to standards as defined in the data dictionary.

Once the data dictionary is completed, a test plan should be developed to ensure that the system implementation supports the data dictionary. This should include sampling data inputs and outputs for conformance, validity, and reliability. This process should also verify interoperability of systems.

13. Provide ongoing education and training for all staff as appropriate to their use of data elements and their definitions.

To ensure consistency of understanding, application, and use of data, it is imperative to provide ongoing education in those definitions. New employee orientation should routinely include exposure to the concepts expressed in the data dictionary.

14. Assess the extent to which the use of the agreed-upon data elements supply consistency of information sharing and avoid duplication.

Ensure simultaneous adoption of new knowledge developed through research and changing terminologies reflective of changes in clinical practice. Specific stakeholders external to most end-user organizations that should be involved in the development and modification of data elements that affect clinical care include all American Board of Medical Specialty recognized specialty societies (e.g., American Academy of Pediatrics and the American Academy of Family Physicians). This evaluation and modification process should be ongoing and involve members of the specialty societies at all stages of the process.

Conclusion

The creation and maintenance of the data dictionary is pivotal to the success of an EHR system. Much thought and effort must go into the planning and the maintenance of this foundational information. Collaboration and buy-in by stakeholders across all domains is critical to the success of the EHR implementation. A process for ongoing maintenance and updates as well as version control must be in place. The upfront design must provide room for change, growth, and expansion over time. Organizations should follow established guidelines such as the ISO/IEC 11179 and the geographic code systems where possible to promote interoperability. Normalization of concepts across end users is an ultimate goal, while any variances in business or clinical concepts should be carefully noted. Once the hard work of the build has been completed, the EHR system should be thoroughly tested to ensure it accurately reflects the standards as defined in the data dictionary.

Prepared by

AHIMA e-HIM Workgroup on EHR Data Content
Carol Adam
Dena Barley
Robert Bishop, MBA, PMP
Keith W. Boone
Christine Brooke, RHIA, CHP
Kathy Callan, MA, RHIA
Barbara Demster, MS, RHIA
Kathy Giannangelo, RHIA, CCS

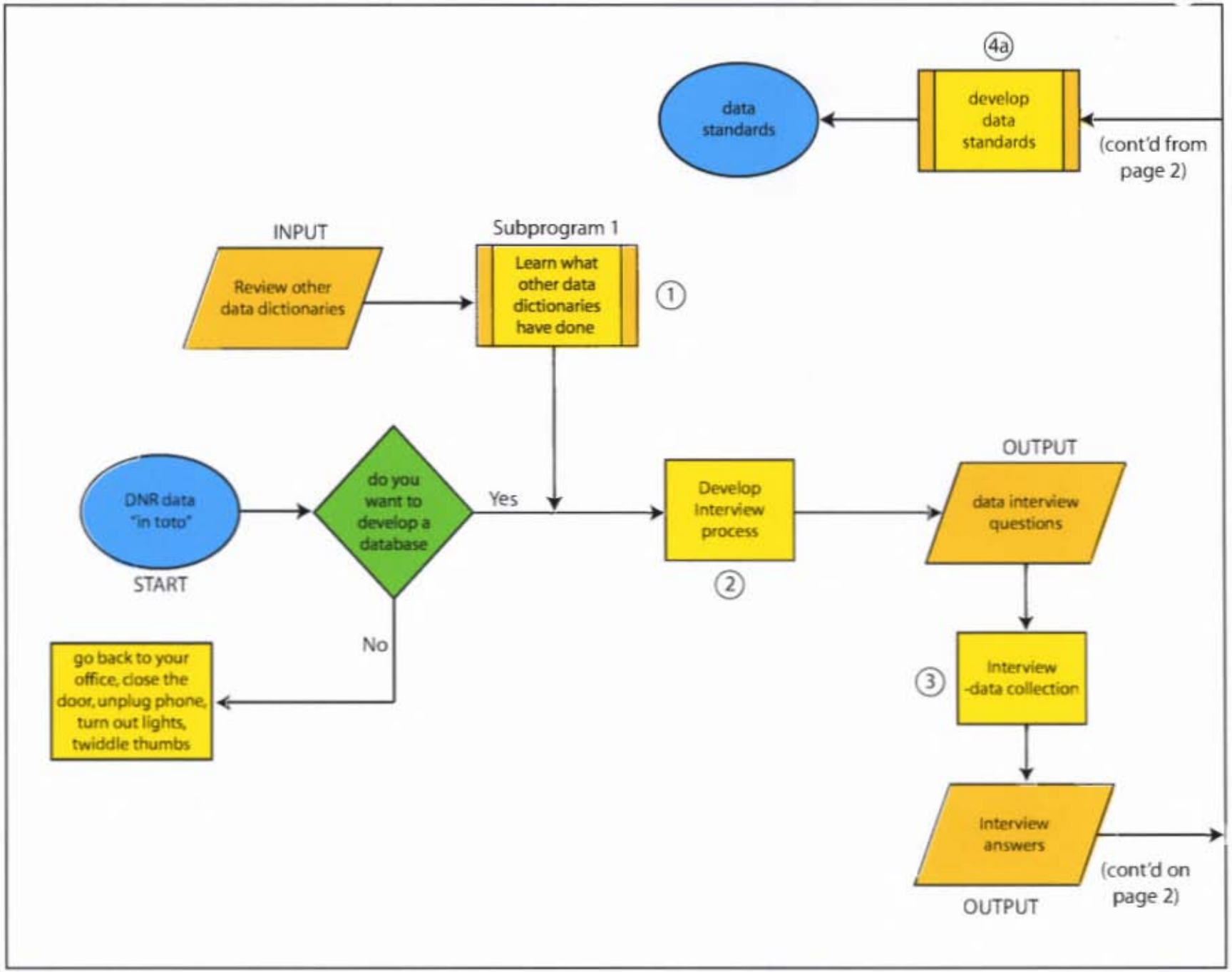
Matthew Greene, RHIA, CCS
Beth Hjort, RHIA, CHPS
Laurie Peters, RHIT, CCS
Christine Rooker, MA, RHIA, CTR
Barbara Samuels, MBA, RHIA
Carol Schuster, RHIA, MSM
Mary H. Stanfill, RHIA, CCS, CCS-P
Dolores Stephens, MS, RHIT
Hao Wang, PhD, MPA
Elmer (Lee) Washington, MD, MPH
Lou Ann Wiedemann, MS, RHIA
Margaret Williams, AM
Carolyn Wilson, MBA, RHIA
Pat Wilson, RT(R), CPC

Article citation:
AHIMA e-HIM Work Group on EHR Data Content. "Guidelines for Developing a Data Dictionary." *Journal of AHIMA* 77, no.2 (February 2006): 64A-D.

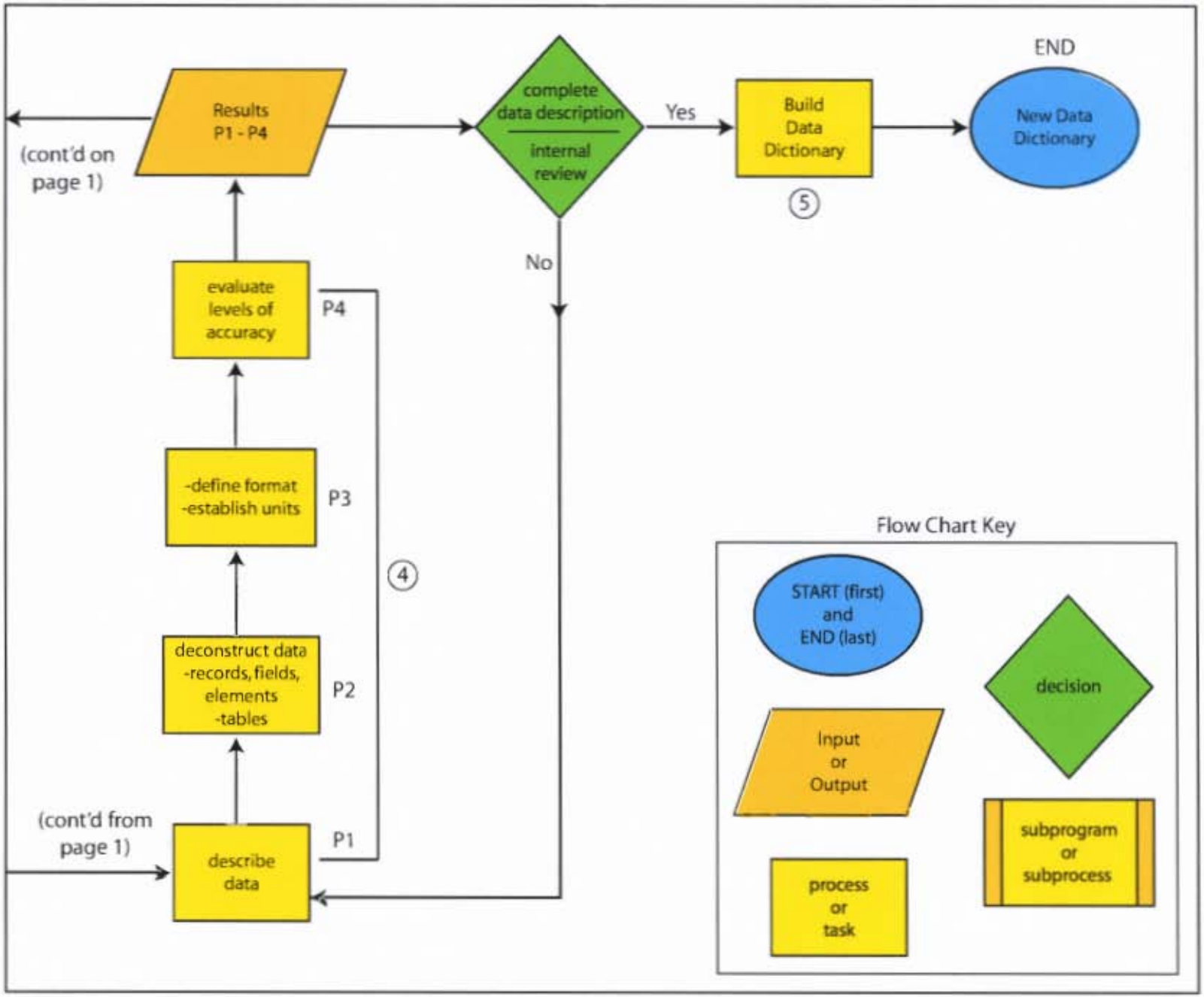
Appendix 4

General Process and Flow Chart for

Developing Data Dictionary



Appendix 4. Flow Chart model for developing data dictionary



Appendix 4. continued

Appendix 4

Notes on generalized process for developing data dictionary correspond to numbered circles on flow chart.

1. Review data dictionaries in other agencies, particularly federal. Most, if not all, agencies in the Department of the Interior have active databases. Along with those databases are usually detailed data dictionaries. These can be a valuable asset as examples of how a dictionary should look. They can provide ideas about how to organize your data, and they can give clarity to problems.
2. Developing interview questions can be reduced to asking who, what, where, when, why type questions. Because this dictionary is being developed by experienced data users, they are supposed to be familiar with the data so that they can ask the "right" questions.
3. The interview process is the main source of information about the data set, and it is the foundation of the dictionary process. An interview can occur in one of two ways. The first is through self-examination by the person building the data dictionary. Although this reduces problems of miscommunication, it has a drawback of possibly not capturing all the information needed. In this case, outside reviews are necessary to ensure a thorough accounting of all the data. The second interview method is the traditional face-to-face conversation with a data handler. In this situation, it is crucial to question all phases of the data, which may require interviewing more than one person. Phases of data include data collection and recording in the field, transfer of data from field to office, data handling and storage in the office, and end uses of data sets.
4. The results from data interviews are put through a series of processes (P1 – P4) that essentially deconstruct (in the non-literary criticism sense) or break down the data into its basic elements: records, fields, and data elements. The data components are given proper names, redundancies are merged or eliminated, definitions are proposed, formats are formalized, units of measurement are agreed on, descriptions of accuracy requirements are made, and detailed descriptions of the data are collected, stored, and utilized are developed. This also results in a subprogram to develop data standards (4a), which are a ancillary result of this project.
5. The data dictionary results from the thorough and accurate description of the collected inputs and outputs. Earlier internal reviews by data users will help refine the final product, which may have several presentation formats. Because it is called a dictionary, it should have the look and feel of a reference source. There is a degree of authority

and permanence given to this document that is almost immutable, so care must be taken when finalizing the product. Once done, the dictionary is essentially cast in stone and revisions to existing entries should be made difficult to do. The only acceptable revisions are additions of new entries.

The organization of the dictionary, however, is unlike regular dictionaries that use alphabetization. Data dictionaries are often organized around data tables. The order of entries can be a map for how the data is collected, or it could be an indication of how significant the data is. The first entries of a data table either being the first information collected at a site or the most important.

Appendix 5

Email Correspondence Concerning

Future Database Development Plans at SCGS

From: Scott Howard
Sent: Thursday, January 31, 2008 4:15 PM
To: Erin Hudson
Subject: FW: what next?

From: Jim Scurry
Sent: Tuesday, December 18, 2007 7:10 PM
To: Scott Howard
Subject: RE: what next?

Scott,
Sorry it has taken so long to get back with you on this. As of today, I believe that the LIDAR project agreement is final, the EDMS is ready for implementation in early January and the Board Room has its overhead projector and screen. All is well with the world.

My thoughts are that our next steps are to:

1 - Have Technology Development Review your design document to gain a general knowledge of the data and processes
2 - Schedule a time to discuss the data and database design parameters & requirements
3 - Make any necessary modifications to the design parameters
4 - Provide metadata, standardized templates and other items that you use to develop the data
5 – Technology allocates staff to integrate/develop the templates or data framework in ArcSDE/Oracle
6 – Identify & develop tools needed by your staff to input, access and maintain the data
7 – Import the data into the integrated DNR database
8 – Establish a regular schedule for data updates/transfer
9 – Routinely revisit core database maintenance issues and status

Off the top of my head these are the primary issues. I may think of others later but am ready to discuss these whenever we can get our schedules in sync.

Thanks for your patience,

Jim

James D. Scurry, Ph.D.
Technology Development Program Director

S.C. Dept. of Natural Resources
Technology Development Program
1000 Assembly Street, Suite 134
Columbia, SC 29201

803-734-9494 (Office)
803-873-1903 (Cell)
803-734-7001 (Fax)

From: Scott Howard
Sent: Monday, November 26, 2007 9:03 AM
To: Jim Scurry
Subject: what next?

Hi Jim,

Holly and Peihua are running a test of the NAS system this week. They'll be out here Thursday to see it in action. Keep you posted.

But... I'm finishing up the CPM project, data dictionary, ideas of data models, and an outline of a general process for developing data dictionaries. My one burning question is what next for the Geological Survey? How do we get from the data dictionary and model to designing and programming a database? What's it going to take, and how can we make sure the momentum doesn't stop? I've got a plan for populating the database with data. I've worked out a timetable and a budget. I may even have most of the needed money. Oh great wizard, what do we do?

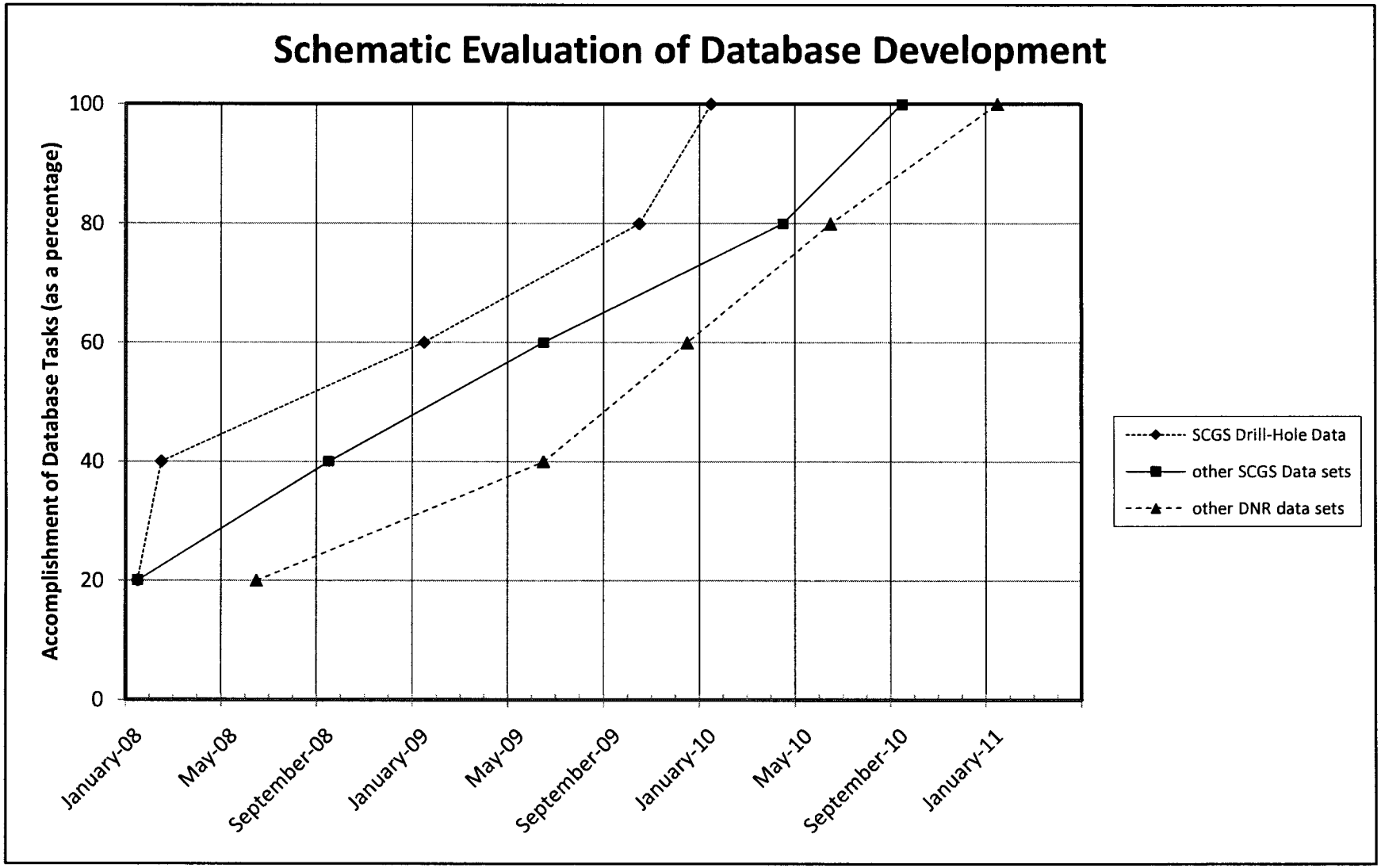
curiously,

Scott

Appendix 6

Benchmarking Database Development:

Examples and Concepts



Appendix 6

Project Measurements

Percent completion	Accomplishment
20	Make commitment to develop database. Although seemingly simple to do, the recording of the start date is important for future benchmarking.
40	Complete data dictionary.
60	Develop Database program. Project is turned over to TD and computer programmers.
80	Programming complete, start data entry, debug program.
100	More than ½ data holdings entered into database. Using database to accomplish job functions.

Chart Assumptions

1. Accomplishment starts at 20 percent for just wanting to develop a database. The commitment to this work is being recognized, but it also marks the start date of the project.
2. Estimate of percentage of project completed is arbitrary and does not reflect a linear or even distribution of work effort or time investment.
3. Target dates are also arbitrary, and they are generously on the high side (over estimations). Better estimates could be developed with discussions of all parties involved.
4. Time estimates (good or bad) will allow post-project statistics to be developed about effectiveness and efficiency of project.